THE EFFECT OF MISSING DATA AND OF TWO SOURCES OF CHARACTER VALUES ON A PHENETIC STUDY OF THE WILLOWS OF CALIFORNIA

THEODORE J. CROVELLO

Numerical taxonomy is still in a phase of expansion and evaluation. New methods continue to appear and more and more taxonomists are using it to estimate phenetic relationships among plant taxa. Gilmartin (1967) provided an extensive bibliography of recent numerical taxonomic applications in botany. Sokal and Sneath (1963) review earlier works.

Concurrent with the development of numerical taxonomy, the use of computers in biological collections and floras has increased. Crovello (1967) lists the uses of electronic data processing in biological collections. Some uses do not involve measurements of characters on specimens, but for other purposes the measurements given in floras would be very helpful. Before information in floras can be employed, the reliability of the measurements provided in a floristic treatment and the problem of missing data must be investigated. By reliability of measurements I mean how accurate they are in estimating the parametric values (e.g., the mean and standard deviation) of a certain character in a given taxon. The problem of missing data arises because many floras are written without the prior preparation of a complete taxon by character table. Many floristic taxonomists include values of a character only in those taxa where the characters are diagnostic of it. This is brought about partly because preserved specimens rarely have all stages of the life cycle present.

The purpose of the present study is twofold: 1, to test the effect of missing data on a phenetic study of a group of plants, and 2, to ascertain the reliability of two sources of information about characters in taxospecies, one arising from a floristic treatment (Munz, 1959) and the other from a study of monographic proportions (Crovello, 1966). The latter is assumed to involve more measurements per character per taxospecies than the former. The results should be of value in estimating the reliability of information from floristic studies for estimation of phenetic relationships. This is especially timely in view of the proposed Flora North Amerca Project. A natural byproduct will be further comprehension of the pattern of variation *among* the willows of California. By this I mean the phenetic relationships among (not *within*) the taxospecies in the context of the characters used. For example, Fig. 1 indicates relationships among the willows based on 43 characters. This figure is a reflection of the pattern of variation among the taxospecies based on these 43 characters.

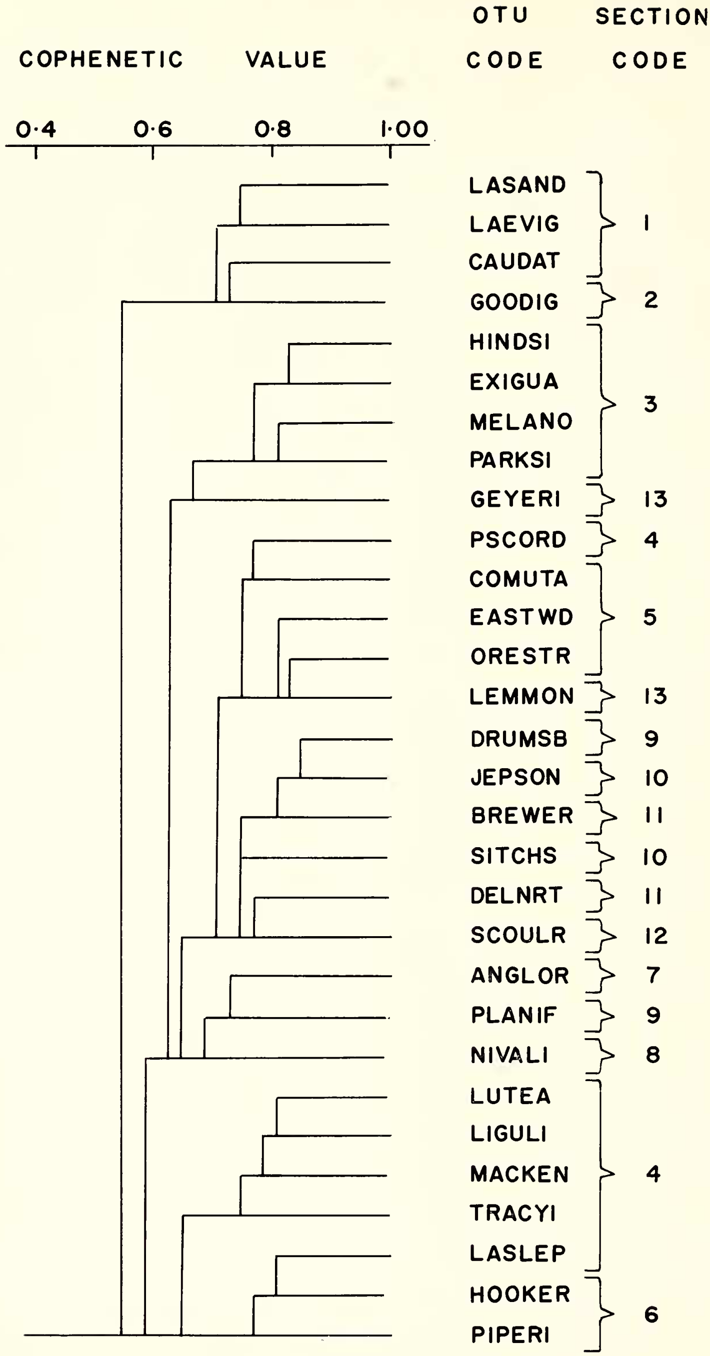


FIG. 1. Phenogram of analysis using data from Crovello (1966) for 43 characters.

Materials and Methods

In his floristic treatment of *Salix,* Munz (1959) used 57 morphological characters, but eight were invariant (did not distinguish any willow species) and were omitted. Of the remaining 49 characters, Crovello (1966) could not obtain information on six of them, so both studies had only 43 characters in common.

Six analyses were made. These differed only in the input data that were used in each. They are the following: 1, data from Crovello (1966) on the 43 available characters of the 57 characters used by Munz (1959); 2, data from Crovello (1966) as the previous analysis, but with the same pattern of missing information as found in Munz (1959); 3, data from Munz (1959) on 43 characters that Crovello (1966) also scored; 4, data from Munz (1959) on 49 of the 57 characters used by him; 5, data from Crovello (1966) on 131 characters; 6, data from Crovello (1966) on 202 characters. Analyses 5 and 6 were included to serve as standards with which to compare the results of the first four analyses. Two standards are considered more desirable, since no one result that uses a large number of characters can be taken as depicting the true, overall, phenetic relationships better than another analysis using many characters. With two "standards" before him, the reader has some idea of the variability of results even when information is available on over 100 characters. Table 1 lists the characters employed in the first four analyses. Crovello (1966) gives the characters used in the last two analyses. In it he treated floral characters common to both sexes, e.g., ament length, as one character, but in the present study these are treated as two characters.

For analyses 3 and 4, the information contained in the treatment of the genus by Munz (1959) was used exclusively. This includes the information located both in the keys and in the description of each species. The information under each variety was incorporated into the description of its species. The other four analyses used information gathered by Crovello (1966). He used herbarium specimens to reinforce his personal collections. From 7 to 15 plants were chosen for each of the 31 taxospecies recognized by Munz (1959) as native to California. Most plants were represented by several herbarium sheets. Crovello (1966) provides a list of specimens used. This list and copies of the taxospecies by character tables, or Basic Data Matrices (BDM's) are on file on punched cards in The Herbarium, University of Notre Dame.

Crovello (1966) concluded that *Salix coulteri* is a synonym of *Salix sitchensis*. As a result, the number of taxospecies analyzed in the present study is 30, one less than the number recognized by Munz. Table 2 lists the 30 taxospecies and their codes. They are grouped into sections according to the ideas of Schneider (1921).

There exists no one method of numerical taxonomy. The present study used only one because we are interested in the effect of different sources of data and not in the effect of different taximetric methods. Sokal and Sneath (1963) discuss a number of different methods.

TABLE 1. CHARACTERS USED IN THE PRESENT STUDY.
THE FIRST 43 WERE USED BY CROVELLO.

1. Plant habit
2. Plant height
3. Last year's twig color
4. This year's twig color
5. Stipules present or absent
6. Stipule length
7. Stipule shape
8. Stipule margin
9. Petiole length
10. Petiole glandular
11. Blade length
12. Blade shape
13. Blade margin entire
14. Blade margin glandular
15. Blade margin revolute
16. Blade base shape
17. Blade apex shape
18. Blade veins prominent below
19. Blade abaxial side glaucous
20. Blade adaxial side lustre
21. Blade width
22. Female ament length
23. Male ament length
24. Female peduncle length
25. Female peduncle leaf number
26. Female ament dense or lax
27. Female floral scale length
28. Female floral scale shape
29. Female floral scale color
30. Stigma lobe length
31. Style length
32. Capsule length
33. Capsule pedicel length
34. Stamen number
35. Stamen filament divided
36. Stamen filament pubescent
37. Presence or absence of pubescence on last year's trig
38. Presence or absence of pubescence on this year's twig
39. Presence or absence of pubescence on abaxial leaf surface
40. Presence or absence of pubescence on adaxial leaf surface
41. Presence or absence of pubescence on adaxial side of floral scale
42. Presence or absence of pubescence on adaxial side of floral scale
43. Presence or absence of pubescence on capsule surface
44. Bark texture
45. Bark color
46. Blade color
47. Time of flowering compared to time of leaf break
48. Ovary shape
49. Flowering period

The procedure used here was the same for each of the six analyses. For each analysis the raw data appears in the form of a taxospecies by character table. or Basic Data Matrix (BDM). Each character was transformed by condensation to remove the effect of weighting due to measurement of different characters in different units. For example, leaf length was measured in millimeters and leaf base shape was measured in angles. To give each character equal weight, each value of a certain character in the tables was condensed, i.e., the value of a character in a certain taxospecies was replaced by a value X_{ei},

$$X_{cj} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

where X_{cj} is the condensed value of character X in taxospecies j, X_j is the original value of character X in taxospecies j and X_{min} and X_{max} are the minimum and maximum observed values of character X in the BDM.

Next, I calculated the similarity between each pair of Operational Taxonomic Units (OTU's), here the taxospecies. I used a modification of the distance coefficient introduced by Sokal (1961). Call this modifi-

MADROÑO

A WEST AMERICAN JOURNAL OF BOTANY

VOLUME XIX

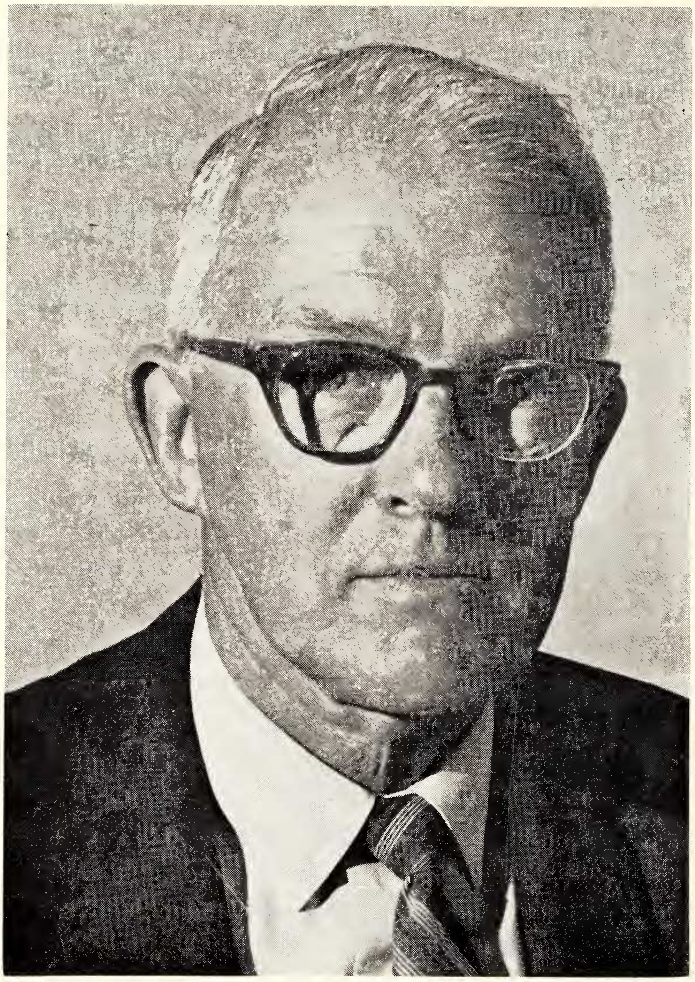
1967–1968

Published quarterly by the

California Botanical Society, Inc.

Life Sciences Building, University of California, Berkeley 94720

Printed by Gillick Printing, Inc., Berkeley, California 94710



To John Thomas Howell, indefatigable collector and student of the plants of California and the Galápagos Islands, and long time member of the California Botanical Society, this volume of Madroño is dedicated.

Tom Howell was born in Merced where he lived until he graduated from high school. It was as a student at the University of California in Berkeley, where he came under the influence of Professor W. L. Jepson, that he developed his interest in botany. Following his university graduation and subsequent completion of a master's degree, he went to southern California, where he spent the first two years of his professional life. In 1929 at the invitation of Miss Alice Eastwood he came to the California Academy of Sciences in San Francisco where he has remained.

As a collector of botanical specimens, Tom Howell has few equals, and his collection numbers have surpassed 40,000. Moreover, his collections have never lain idle to collect dust but he has determined them and incorporated them in the Herbarium of the Academy. Although most of his collecting has been done in California and the Galápagos Islands he has made important collections in several western states. In addition to his own collections which have added so extensively to the Academy's herbarium he has encouraged many others, both amateur and professional, to make collections which also have come to the Academy.

His numerous botanical papers have contributed greatly to our knowledge of the plants of the areas where he has collected and particularly his floras of several regions in northern and central California have contributed to a greater botanical understanding of local areas.

Leaflets of Western Botany, which he and Miss Eastwood began in 1932 and which he edited and published until the end of 1966, provided a vehicle for significant contributions to the botany of the western United States.

His many years of field work, his keen observations, his great enthusiasm and unlimited energy, have brought to him an unsurpassed knowledge of California's native and weedy plants. And he is always a source of botanical information, willingly, thoughtfully and kindly given, to all who seek him out.

CONTENTS

STATEMENT OF OWNERSHIP, MANAGEMENT, AND CIRCULATION

(Act of Oct. 23, 1962; Section 4369, Title 39, United States Code)

MADROÑO, A West American Journal of Botany, is published quarterly at Berkeley, California.

The publisher is the California Botanical Society, Inc., Life Sciences Building, University of California, Berkeley, California 94720.

The editor is John H. Thomas, Division of Systematic Biology, Stanford University, Stanford, California 94305.

The owner is the California Botanical Society, Inc., Life Sciences Building, University of California, Berkeley, California 94720. There are no bondholders, mortgagees, or other security holders.

The average number of copies distributed of each issue during the preceding 12 months is 650; the numbers of copies of the single issue closest to the filing date is 650.

I certify that the statements made by me above are correct and complete.

JOHN H. THOMAS, Editor

November 15, 1968

TABLE 2. THE 30 TAXOSPECIES OF SALIX IN CALIFORNIA RECOGNIZED BY CROVELLO (1966). ARRANGEMENT BY SECTIONS FOLLOWS SCHNEIDER (1921)[1].

Section Code	Taxospecies Symbol	Taxospecies
1. *Pentrandrae* Dumort.	LASAND	*S. lasiandra* Benth.
	CAUDAT	*S. caudata* (Nutt.) Hell.
	LAEVIG	*S. laevigata* Bebb.
2. *Nigrae* Loudon	GOODIG	*S. gooddingii* Ball
3. *Longifoliae* Anderss.	HINDSI	*S. hindsiana* Benth.
	EXIGUA	*S. exigua* Nutt.
	MELANO	*S. melanopsis* Nutt.
	PARKSI	*S. parksiana* Ball
4. *Cordatae* Barr.	LUTEA	*S. lutea* Nutt.
	LIGULI	S. *ligulifolia* (Ball) Ball
	MACKEN	S. *mackenziana* (Hook.) Barr.
	PSCORD	*S. pseudocordata* Anderss.
	LASLEP	*S. lasiolepis* Benth.
	TRACYI	*S. tracyi* Ball
5. *Andenophyllae* Schneid.	COMUTA	S. *commutata* Bebb
	EASTWD	*S. eastwoodiae* Ckll.
	ORESTR	*S. orestera* Schneid.
6. *Chrysanthae* Koch	PIPERI	*S. piperi* Bebb
	HOOKER	*S. hookeriana* Barr.
7. *Ovalifoliae* Rydb.	ANGLOR	*S. anglorum* Cham. var. *antiplasta* Schneid.
8. *Reticulatae* Fries	NIVALI	*S. nivalis* Hook.
9. *Phylicifoliae* Dumort.	PLANIF	*S. planifolia* Pursh var. *monica* (Bebb) Schneid.
	DRUMSB	*S. drummondiana* Barr. var. *subcoerulea* (Piper) Ball
10. *Sitchenses* Bebb	SITCHS	*S. sitchensis* Sans.
	JEPSON	*S. jepsonii* Schneid.
11. *Brewerianae* Schneid.	BREWER	S. *breweri* Bebb
	DELNRT	*S. delnortensis* Schneid.
12. *Discolores* Barr.	SCOULR	*S. scouleriana* Barr.
13. *Fulvae* Barr.	LEMMON	*S. lemmonii* Bebb
	GEYERI	*S. geyeriana* Anders.

[1] The only exception to Schneider's assignments is *S. jepsonii.* He placed it in section *Phylicifoliae.*

cation the similarity coefficient, s_{jk}. Then,

$$s_{jk} = 1 - \left[\frac{\sum_{i=1}^{n_{jk}} (X_{ij} - X_{ik})^2}{n_{jk}} \right]^{1/2}$$

where X_{ij} and X_{ik} are the values of character i in OTU's j and k, respectively, and n_{jk} is the total number of characters used in the particular comparison. If there were no missing data, i.e., if the OTU by OTU relevance (Sokal and Sneath, 1963) were always 1.0, then n_{jk} would be the same for all combinations of pairs of OTU's.

TABLE 3. SUMMARY STATISTICS OF THE SIX ANALYSES.

Analysis number and descripion	Similarity coefficient		OTU by OTU relevance		Cophenetic correlation coefficient
	$\overline{Y}$	s	$\overline{Y}$	s	
1. Crovello 43	.593	.086	.972	.037	.746
2. Crovello 43 with Munz's missing data pattern	.578	.098	.482	.082	.678
3. Munz 43	.590	.085	.495	.079	.741
4. Munz 49	.587	.082	.483	.075	.719
5. Crovello 131	.617	.077	.908	.032	.858
6. Crovello 202	.615	.072	.829	.080	.871

Each set of similarity coefficients from one BDM forms a 30 by 30 OTU by OTU table, or Basic Similarity Matrix (BSM). Each cell in it indicates how similar two taxospecies are in the context of the characters used and based on the source of the data for the BDM being analyzed. The BSM was then used to group OTU's by the unweighted pair-group method (Sokal and Sneath, 1963). This results in a phenogram, a hierarchic presentation of phenetic relationships among the OTU's, *in the context of the characters analyzed.*

RESULTS

Figures 1 to 4 and Tables 3 and 4 present the results of the four analyses. Figures 5 and 6 are standards with which to compare the phenograms of the analyses using Munz's characters. The last two represent the maximum information on California willows available to the author at the present time. Figure 5 is based on the morphological characters studied by Crovello (1966) but without the six (or seven) unit pubescence characters scored for each organ. Figure 6 includes the 62 unit pubescence characters. *Two* standards were used: 1, to increase comprehension of the results of the present paper; and 2, to emphasize that in numerical taxonomy any one result is not the ultimate truth. For ease of presentation of results, we shall compare Figs. 1 through 6 with the latest nonnumerical monograph, which I summarize in Table 2.

Figure 1 gives the analysis of 43 characters using full data from Crovello (1966) which produced seven clusters. Beginning at the top, the first cluster contains four taxospecies that are the only representatives in California of the subgenus *Pleiandrae*. The next four OTP's are the representatives of section *Longifoliae* appearing in California. GEYERI then joins this cluster. The next cluster of five taxospecies includes members of sections *Cordatae, Adenophyllae* and *Fulvae,* while the subsequent cluster includes members of sections *Phylicifoliae, Sitchenses, Brewerianae* and *Discolores* in a mixed pattern. This is followed by a cluster consisting of three high-altitude willows, ANGLOR, PLANIF and NIVALI. The next to last cluster includes four members of section *Cordatae,* while the last cluster contains LASLEP, a polymorphic member of sec-

TABLE 4. CORRELATION BETWEEN ALL PAIRS OF THE SIX SIMILARITY MATRICES (LOWER LEFT), AND CORRELATION BETWEEN ALL PAIRS OF THE SIX PHENOGRAMS (UPPER RIGHT). IN ALL CASES n = 435.

Analysis number and description	Analysis number					
	1	2	3	4	5	6
1. Crovello 43	1.000	.852	.606	.643	.715	.677
2. Crovello 43 with Munz's mising data pattern	.887	1.000	.455	.437	.460	.485
3. Munz 43	.583	.450	1.000	.947	.658	.573
4. Munz 49	.604	.460	.974	1.000	.739	.653
5. Crovello 131	.835	.669	.655	.670	1.000	.889
6. Crovello 202	.733	.585	.595	.614	.889	1.000

tion *Cordatae,* and the two taxospecies of section *Chrysanthae* that appear in California.

Turning to Fig. 2, which is based on Crovello's data but uses the pattern of missing data present in the Munz data, seven clusters also are seen here. The first is similar to that of Fig. 1. The next two clusters of three and four taxospecies resemble two from Fig. 1, but TRACYI is out of place. The next cluster includes section *Longifoliae,* but here GEYERI has split the four members of that section. The next cluster contains six taxospecies. Except for ANGLOR, it is similar to a cluster in Fig. 1. The final two clusters also have their counterparts in Fig. 1, with the exception of ANGLOR mentioned above.

Figure 3 is based on Munz's data on the 43 characters that are comparable to Crovello's. The first cluster of four taxospecies is the same as in previous figures. But then HINDSI appears as deviant from all other OTU's. The other three members of section *Longifoliae* are far removed from it. The next two clusters, the first with six taxospecies, bring together quite different OTU's as suggested by conventional taxonomy and by Figs. 1 and 2. ANGLOR is a dwarf alpine form, whereas SCOULR is a polymorphic shrub or small tree more common at lower altitudes. By inspection of the rest of Fig. 3 the reader can ascertain the similarities and differences of it when compared to the previous figures.

Figure 4 is based on the 49 characters used by Munz. Seven more or less distinct clusters emerge. The four OTU's of section *Longifoliae* are closer together now than in Fig. 3, but as in Fig. 2, GEYERI splits them. Here EASTWD also is among this cluster. Note at the bottom of Fig. 4 that ANGLOR is still grouped with SCOULR.

Figure 5 consists of eight clusters, the first three of which agree exactly with Fig. 1. The remainder of Figs. 1 to 4 is in less agreement with Fig. 5 and Fig. 6. Note, however, that Figs. 5 and 6 are not identical.

Table 3 gives summary statistics for the six analyses. Columns 1 and 2 list the mean and standard deviation of each of the Basic Similarity